

# Venues and Crimes in Atlanta

## Introduction

My wife and I are currently living in Long Beach, CA and she has a job opportunity in Atlanta, GA. Before we just look for places to live, I would to do some analysis about Atlanta, regarding which neighborhoods are safer.

## Problem

I want to see if there is a correlation between top venue categories and the number of crimes committed. For example, if there's a neighborhood that has a higher number of bars/restaurants, will there be a higher crime rate at that particular neighborhood? I will use K-Means algorithm to cluster certain neighborhoods together and compare similar characteristics that each neighborhood obtains to help make a decision. I will also do some Exploratory Data Analysis to identify any trends of crimes, venues or anything else that comes to mind.

## Data

The data I will gather to analyze for this project are:

1. Crime data appertaining to Atlanta from January 2018 to December 2018.
2. Geographical data of neighborhoods from Foursquare.

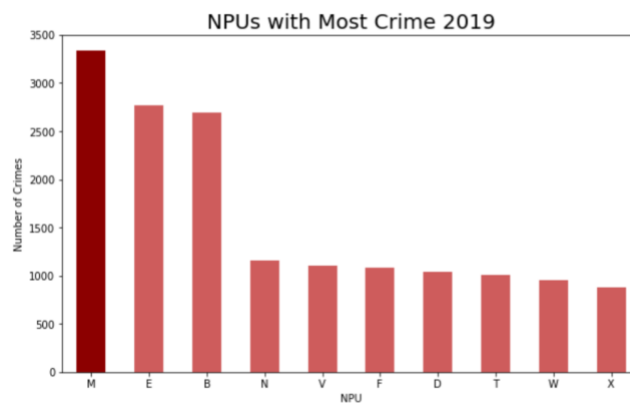
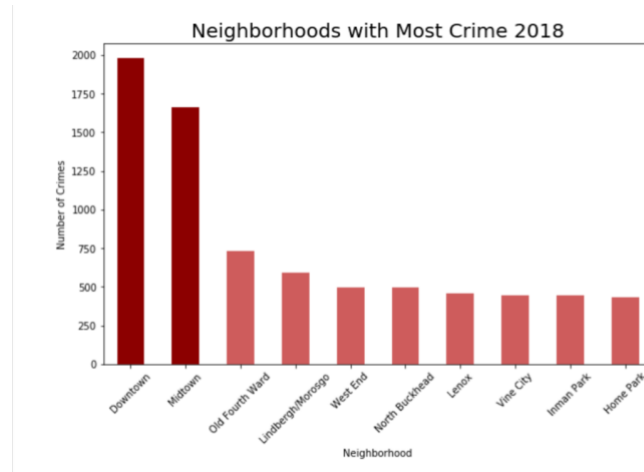
Here are the links of the data I am using:

- [https://en.wikipedia.org/wiki/Neighborhood\\_planning\\_unit](https://en.wikipedia.org/wiki/Neighborhood_planning_unit)
- <http://opendata.atlantapd.org/Crimedata/Default.aspx>

It is worth noting that I manually inputted the NPU data in an excel file. With those two links, I will scrape data from those websites so I can build two data frames. I will build a data frame for crime and one for venues which will include latitude and longitude. The crime dataset for Atlanta (2019), has 12 files (1 for each month). I concatenated 12 csv files and formed it into one data frame. In addition, I will proceed with any adjustments necessary to clean the data. It will consist of removing rows, with missing information, removing duplicate rows, removing unnecessary columns and other adjustments of the sort.

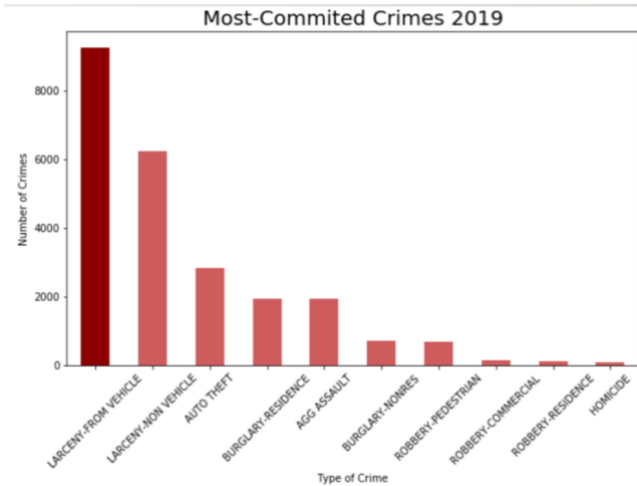
## Exploratory Data Analysis

The bar chart below represents the top ten neighborhoods of their total number of crimes (overall) throughout the year. According to the crime data, Downtown and Midtown are the top two neighborhoods with the highest crime rate. Old Fourth Ward comes in third place with a vast drop.

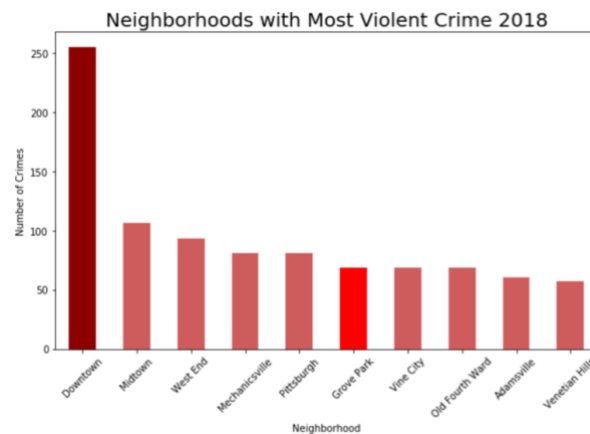


The bar chart above represents the top ten NPU's with the highest overall crime rate. The top three NPU's alone make up almost 33% of the overall crime in Atlanta.

The bar chart below represents the top ten crimes such as assault, robberies and so on that were committed in 2019:

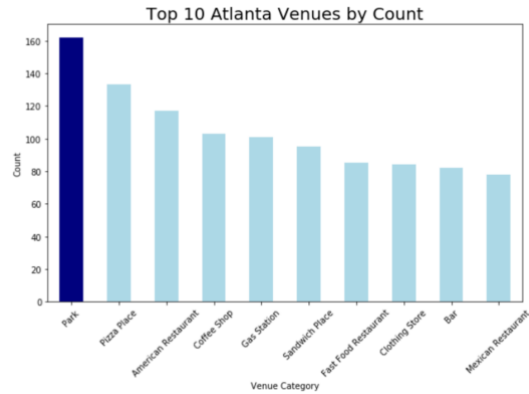


As we can see, larceny from vehicle and from non-vehicle is the most common crime in Atlanta in 2019.



The bar chart above describes the relationship with the number of violent crimes such as assault, homicide, etc. in each neighborhood. Downtown is the neighborhood with the highest violent crime rate and remember, it was also the highest in overall crime rate. It is interesting to point out that Grove Park made the list in the top ten violent crimes when that neighborhood was not in the first bar chart (neighborhoods with overall crime rate).

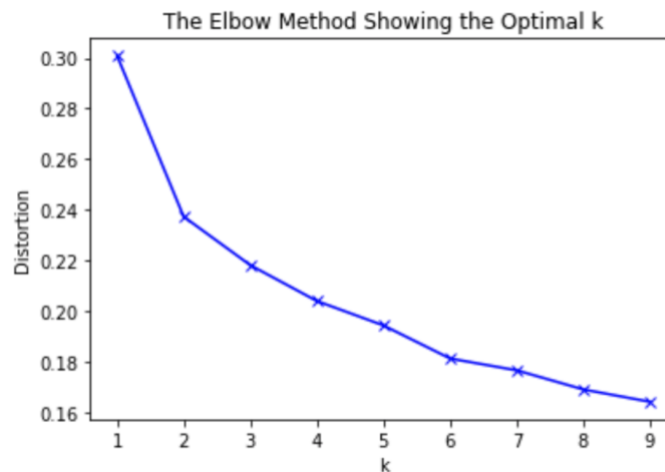
The bar chart below represents the top ten venues in Atlanta. Venues can be anything like parks, schools, restaurants, museums and so on.



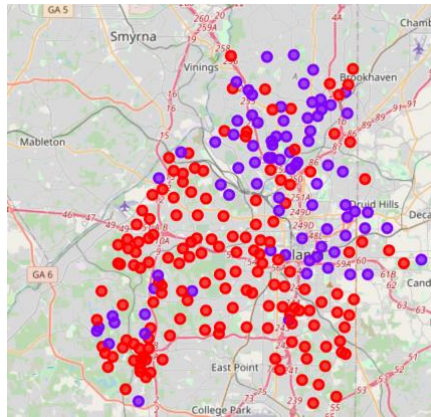
The top venue in Atlanta is parks. The disparity between venues is not much but it is still important to visualize which venues are the most popular.

## Methodology

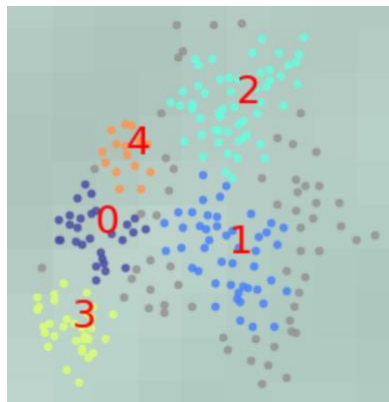
I wanted to cluster neighborhoods based on location and crime rate. I also clustered neighborhoods based on a specific type of crime. I used K-Means and DBSCAN Clustering methods. I used several values for 'k' to give me the optimal model. According to the table below, and based on the Elbow Method, the optimal model for  $k$  is  $k = 2$ .



When I chose to cluster neighborhoods using only two clusters, I used the Folium package in Python to give me a visualization.



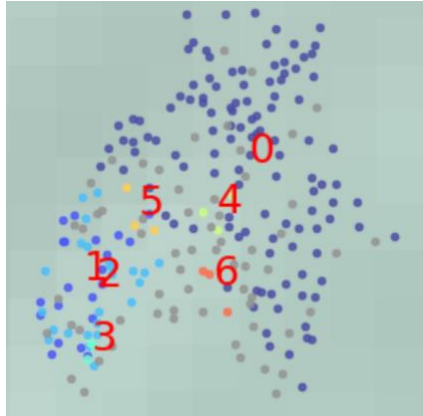
The next visualization of clusters represents the clusters of each neighborhood based on location and overall crime rate in 2019.



The average number of total crimes per cluster are:

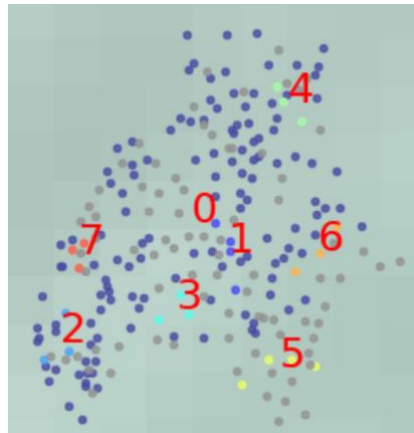
- Cluster 0: 34
- Cluster 1: 172
- Cluster 2: 91
- Cluster 3: 33
- Cluster 4: 47

The next visualization of clusters represents the clusters of each neighborhood based on location and violent crime rate in 2019.



Cluster 0, Avg Number of Homicide/Agg Assault Crimes per Neighborhood: 128.88793103448276  
 Cluster 1, Avg Number of Homicide/Agg Assault Crimes per Neighborhood: 7.142857142857143  
 Cluster 2, Avg Number of Homicide/Agg Assault Crimes per Neighborhood: 15.166666666666666  
 Cluster 3, Avg Number of Homicide/Agg Assault Crimes per Neighborhood: 6.666666666666667  
 Cluster 4, Avg Number of Homicide/Agg Assault Crimes per Neighborhood: 151.0  
 Cluster 5, Avg Number of Homicide/Agg Assault Crimes per Neighborhood: 154.0  
 Cluster 6, Avg Number of Homicide/Agg Assault Crimes per Neighborhood: 218.66666666666666

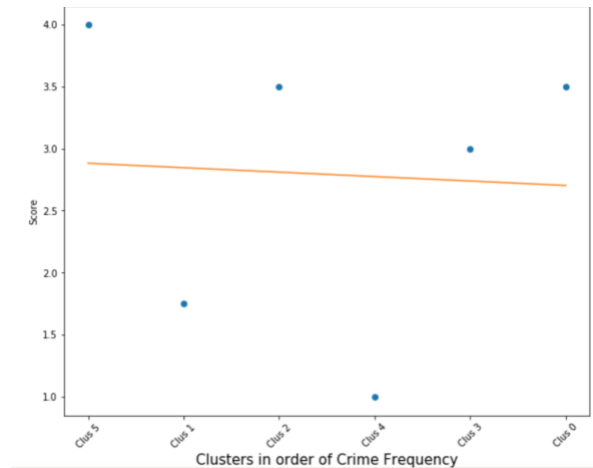
The next visualization of clusters represents the clusters of each neighborhood based on location and robberies in 2019.



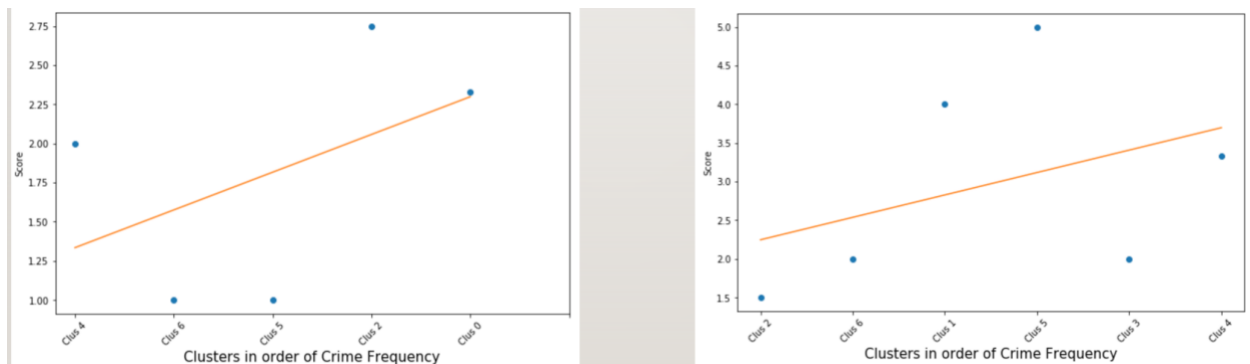
Cluster 0, Avg Number of Robbery Crimes per Neighborhood: 78.6923076923077  
 Cluster 1, Avg Number of Robbery Crimes per Neighborhood: 204.5  
 Cluster 2, Avg Number of Robbery Crimes per Neighborhood: 25.0  
 Cluster 3, Avg Number of Robbery Crimes per Neighborhood: 133.66666666666666  
 Cluster 4, Avg Number of Robbery Crimes per Neighborhood: 290.3333333333333  
 Cluster 5, Avg Number of Robbery Crimes per Neighborhood: 56.75  
 Cluster 6, Avg Number of Robbery Crimes per Neighborhood: 237.0  
 Cluster 7, Avg Number of Robbery Crimes per Neighborhood: 17.333333333333332

## Results

The objective of this analysis is to determine a correlation with crime and certain type of venues. I used three different plots, each with different kinds of venues to see if there is a correlation.



The correlation between overall crimes and venues are slightly negative which is surprising. I did not expect to see this result.



The plot on the left is the correlation between neighborhoods and violent crimes. The plot on the right represents the correlation between neighborhoods and robberies. As you can see, the plot on the left is a stronger, positive correlation and the plot on the right appears to show a positive correlation but not as strong as the correlation between neighborhoods and violent crimes.

## **Conclusion**

One of my goals was to search for correlations venues and crime. For example, if there is a neighborhood that had more liquor stores or bars, then that can be an indication of resulting in a higher crime rate. It is also important to consider that there may be other variables or factors that can contribute to crime rate. I used crime rank and gave an average score for each cluster for and used those to plot my points in my scatterplots. For example, a cluster with a higher frequency of a particular venue received a score of 1.

Based on these findings, I can conclude that there is a positive correlation between alcohol selling venues (bars, liquor stores) and crime rate. If I were to encounter another dataset that had more variables, then the results of this analysis would have a stronger and more meaningful conclusion. Unfortunately, with the datasets I worked with, I was limited.